사례기반 학습을 이용한 음절기반 한국어 단어 분리 및 범주 결정

김 재 훈[†] · 이 공 주[††]

요 약

한국어는 영어와 같이 공백을 단어의 경계로 사용하지만, 그 단어의 구조는 영어와 다소 차이가 있다. 영어는 일반적으로 공백 사이에 하나의 단어가 포함되나, 한국어는 여러 개의 단어 혹은 형태소가 포함된다. 이런 차이 때문에 일반적으로 한국어에서는 공백을 경계로 이루어진 단어를 어절이라고 한다. 본 논문에서는 하나의 어절 내에 포함된 단어들을 분리하고, 분리된 각 단어의 적절한 범주를 결정하는 방법을 제안한다. 본 논문에서는 사례기반 기계학습 방법을 이용하고 음절 단위로 단어를 분리한다. 사례기반 학습을 위해 사용된 자질집합은 이전 음절, 자신의 음절, 이후의 두 음절, 자신의 음절에 대한 받침 정보, 이전 두 범주 정보이다. 제안된 시스템을 평가하기 위해서 ETRI 말뭉치와 KAIST 말뭉치를 사용하였으며, 두 말뭉치 모두에서 단어 분리의 F 측도가 97% 이상으로 비교적 좋은 성능을 보였다.

Segmenting and Classifying Korean Words based on Syllables Using Instance-Based Learning

Jae-Hoon Kim[†] · Kong Joo Lee[††]

ABSTRACT

Korean delimits words by white-space like English, but words in Korean is a little different in structure from those in English. Words in English generally consist of one word, but those in Korean are composed of one word and/or morpheme or more. Because of this difference, a word between white-spaces is called an Eojeol in Korean. We propose a method for segmenting and classifying Korean words and/or morphemes based on syllables using an instance-based learning. In this paper, elements of feature sets for the instance-based learning are one previous syllable, one current syllable, two next syllables, a final consonant of the current syllable, and two previous categories. Our method shows more than 97% of the F-measure of word segmentation using ETRI corpus and KAIST corpus.

키워드 : 단어 분리(Word segmentation), 사례기반 학습(Instance-based learning)

1. 서 론

한국어는 영어와 같이 공백을 단어의 경계로 사용하지만, 그 단어의 구조는 영어와 다소 차이가 있다. 영어는 일반적으로 공백 사이에 하나의 단어가 포함되나, 한국어는 여러 개의 단어 혹은 형태소가 포함된다. 이런 차이 때문에 일반적으로 한국어에서는 공백을 경계로 이루어진 단어를 **어절**[1] 이라고 한다. 본 논문에서는 하나의 어절 내에 포함된 단어들을 분리하고, 분리된 각 단어의 적절한 범주를 결정하는 방법에 대해서 기술한다.

한국어 처리 분야에서 단어 혹은 형태소의 경계를 찾고 그들의 범주를 결정하기 위해서 주로 형태소 분석기(morphological analyzer)와 품사 부착 시스템(part-of-speech tagger)을 사용한다[2-3]. 이 경우, 정확한 단어 분리 결과를 얻을 수 있으며, 또한 용언의 활용으로 변형된 어간과 어미에 대해서도 원형을 복원할 수 있다. 이 때문에, 형태소 분석기나 품사 부착 시스템은 자연언어 처리 시스템의 기본 시스템으로 사용되고 있다. 그러나 형태소 분석기이나 품사 부착 시스템은 매우 복잡하고, 이들을 구현하기 위해서는 복잡한 언어지식과 방대한 사전정보가 요구된다. 응용분야에 따라서는 형태소 분석이나 품사 부착의 결과로서 모든 품사의 단어를 이용하지 않는다. 예를 들면, 일부의 정보검색 시스템은 주어진 문서에서 명사만을 추출하여 문서를 색인한다. 이 경우, 형태소 분석에서 용언을 분석하기 위한 복잡한 과정[2]이 필요하지 않을 수 있다.

본 논문에서는 형태소 분석기와 품사 부착 시스템과 같은 언어처리 시스템이나 대량의 사전정보를 이용하지 않고

※ 이 논문은 한국과학재단의 해외 post-doc 연구지원비와 2002년도 두뇌한국(BK)21 사업에 의해서 지원되었음.
† 정 회 원 : 한국해양대학교 컴퓨터공학과 교수
†† 정 회 원 : 이화여자대학교 컴퓨터공학과 교수
논문접수 : 2002년 8월 7일, 심사완료 : 2002년 9월 24일

단어를 분리하고, 분리된 단어의 범주를 결정하기 위한 새로운 방법을 제안한다. 이 방법은 품사 부착 말뭉치로부터 단어의 주변 환경을 학습하여 주어진 문장으로부터 단어를 인식하고, 그 단어의 범주를 결정하는 방법이다. 본 논문에서 단어의 경계(word boundary)는 음절로 제한한다. 이 경우, 음운현상이나 용언의 활용으로 단어의 경계가 음절로만 제한되지 않는다는 문제가 발생된다. 예를 들면 용언의 어간 "가"와 어미 "ㄴ다"가 활용되면 "간다"가 되므로 어절 "간다"의 형태소 분석 결과는 "가/동사 + ㄴ다/어미"된다. 따라서 어절 "간다"가 단어로 분리되면 "가 + ㄴ다"가 되므로 음절 "간" 내에 단어의 경계가 존재하게 된다. 이와 같은 현상은 주로 용언에서 발생하며, 용언의 유형에 따라 일정한 규칙을 가지므로 각 유형에 따라 음절기반 단어 분리 경계를 정의한다. 예를 들면, 앞의 예에서 어절 "간다"의 경우는 "간 + 다"로 분리한다. 이와 같이 분리된 결과는 응용분야에 따라서 형태소를 분석하여 사용할 수도 있고, 단어 분리 결과를 그대로 사용할 수도 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 단어 분리 방법론과 본 논문과 관련된 기계학습 방법에 대해서 살펴보고, 3장에서는 음절기반 단어 분리 방법에 대해서 논하고, 4장에서는 제안된 시스템의 성능을 평가한다. 마지막으로 5장에서 결론을 맺고 앞으로의 연구 방향을 제시한다.

2. 관련 연구

2.1 단어 분리

한국어 단어분리에 대해서는 다음절에서 자세히 언급하고, 본 절에서는 일반적인 단어 분리 방법론에 대해서 간략히 소개한다.

음성인식 등의 분야에서 단어 분리[4-5]를 다루지만, 자연언어 처리에서 말하는 단어분리와는 다소 차이가 있기 때문에 본 절에서는 주로 자연언어 처리 분야에서 연구된 단어 분리 방법론에 대해서 살펴볼 것이다. 단어 분리 방법으로 가장 간단한 방법은 n-그램 방법[6]이다. 이 방법은 엄밀히 말하면 단어 분리라기보다는 정보검색의 색인 방법이다. [7]에서는 통계 유한상태 모델을 이용한 단어 분리 모델을 제안했다. 이 방법은 모든 단어를 가중치 유한상태 오토마타(사전 오토마타)로 표현하고, 문장을 하나의 오토마타(문장 오토마타)로 표현한다. 단어 분리는 이 두 오토마타를 합성하여(compose) 최적경로를 찾으므로 이루어진다. 오토마타를 이용할 경우에는 미등록어를 처리하기 어렵지만, 실행속도가 빠르고, 많은 도구가 존재하므로 빠른 시간 내에 시스템을 개발할 수 있다는 장점이 있다. [8]에서는 변형기반 방법을 단어 분리에 적용했다. 이 방법은 단어 분리 환경에 해당하는 변형 규칙 틀(transformation rule template)[3]을 정의하고, 정의된 규칙 틀과 변형기반 학습 알고리즘[9]을 이용해서 자동으로 규칙을 추출한다. 추출된 규칙을 이용해서 주어진 문장에 공백을 넣거나 잘못 넣은 공백을 제거하는 과정을 반복하여, 단어를 분리한다. [10]에서는 정보이론을 단어 분리에 적용하였다. 이 방법의 기본 개념은 음절간의 상호정보(mutual information)와 단어의 정보량(information content) 그리고 사전을 이용해서 단어를 분리한다. 일단 사전을 이용해서 단어로 분리될 수 있는 후보를 선정한다. 그 결과에 중의성이 있을 경우, 단어의 정보량의 변화가 가장 큰 곳을 단어 분리 위치로 결정하는 방법이다. 사전 탐색은 양방향 최장일치를 적용하는데 사전 탐색 동안에 분리될 단어 후보가 제대로 선정되지 못할 경우, 정확하게 단어를 분리할 수 없다는 단점을 가지고 있다. 이런 문제를 보완하기 위해 [11]에서는 상호정보량이 최소인 지점에서 분리하고, 상호정보량이 최대인 지점에서 결합하는 방법을 제안하여 이 문제를 개선하였다. [12]에서는 은닉 마르코프 모델(hidden Markov model)을 단어 분리 모델에 적용하였다. 이 방법은 주어진 문자열에 대해서 공백을 추가하는 방법으로 모델링되었으며, 고차 은닉 마르코프 모델을 사용하며 비교적 좋은 결과를 얻었다.

2.2 한국어 단어 분리 방법론

앞에서 언급했듯이 한국어 단어 분리에 대한 연구는 중국어나 일본어에 비해 활발하게 진행되지 않았다. 한국어에서도 n-그램을 이용한 단어 분리 방법[13]이 연구되었으나, 이 방법은 정보검색의 색인어를 추출하기 위해서 사용된 방법으로 조사나 어미와 같은 기능어를 먼저 분리한 후, 내용어에 대해서는 n-그램을 이용해서 색인어를 추출한다. 앞에서도 언급했듯이 이 방법은 단어 분리라기보다는 색인어 추출 방법의 일종이다.

한국어 단어 분리는 주로 형태소 분석기의 기본적인 기능 중 하나로 간주되었다. 대부분의 형태소 분석에서는 단어 분리를 위해서 주로 사전과 결합규칙(morphotactics)을 이용하며, 형태소 분석기에 따라서 모든 품사의 사전을 이용할 수도 있고, 조사나 어미 사전만 이용하는 경우가 있다. 전자의 경우는 형태소 분석 알고리즘과 밀접하게 관계되어 있어서 단어 분리를 위한 독립적인 모듈을 가지지 않는다. 후자의 경우는 단어 분리를 위해서 독립적인 모듈을 가지며, (복합)어미와 (복합)조사 사전, 결합규칙 그리고 조

1) 연세 한국어 사전[1]에서 어절에 대한 정의를 살펴보면, 어절은 (문장 성분의 최소 단위로 띄어쓰기의 단위가 되는) 문장을 이루는 도막 도막의 성분이다.

2) 용언에 대한 형태소 분석은 활용 처리, 불규칙 처리, 음운현상 처리 등 매우 복잡한 과정을 포함하고 있으며, 실질적으로 형태소 분석 시스템의 프로그램 크기의 1/3 이상을 차지한다.

3) 중국어에서 단어분리를 위한 규칙 틀을 예를 들어 보면 다음과 같다.
AB ⇒ A B ; xB ⇒ x B ; Ay ⇒ A y ; JAB ⇒ JA B ;
여기서 A, B, J는 특정 문자를 표현하고, x, y는 모든 문자가 될 수 있다.

사와 어미의 음절정보를 이용해서 주어진 어절로부터 조사와 어미, 단음절 명사, 보조용언, 접사를 분리한다[14].

한국어에서 단어 분리에 대한 독립적인 연구로는 [15]가 있다. 이 연구는 2-그램의 음절정보를 이용한 단어 인식 모델을 제안하였으나, 엄밀한 의미에서 단어 분리 모델이라기보다는 띄어쓰기 모델이라고 볼 수 있다. 왜냐하면 문장이 입력되면 공백을 제거한 후, 일본어 문장이나 중국어 문장과 같이 공백이 없는 문장에 대해서 공백을 추가하여 단어를 분리한다. 따라서 이 모델은 모든 문장의 입력은 오류를 포함한다고 기본적으로 가정하고 있다. 이 모델의 기본적인 개념은 공백을 포함하는 음절 2-그램과 공백을 포함하지 않는 음절 2-그램의 확률과 단어 2-그램 확률을 이용한 언어모델(language model)을 이용하였다.

2.3 한국어 단어 분리의 필요성

한국어 처리 분야에서 단어 분리에 대한 독립적인 연구는 그다지 활발하지 않았다. 그 이유에 대해서 생각해 보자. 첫째, 중국어와 일본어와는 다르게 한국어는 단어 분리를 위해서 공백을 사용한다. 따라서, 영어의 경우와 같이 단어 분리 문제를 대수롭게 여기지 않았다. 둘째, 음운현상이나 용언의 활용 등으로 단어분리에 대한 기준이 명확하지 않다. 셋째, 단어 분리도 형태소 분석만큼의 중의성을 가지고 있다.

최근에 인터넷의 기술이 발전되고, 대량의 정보를 빠른 시간 내에 처리하기 위해서 한국어 처리 분야에서도 단어 분리에 대한 관심이 차츰 증가되고 있다. 한국어에서 일반적인 단어 분리가 필요한 이유를 살펴보자. 첫째, 일반적으로 형태소 분석기는 미등록어 처리에 많은 문제를 보인다. 대부분의 형태소 분석기들은 오른쪽에 있는 기능어를 분석한 후, 결합규칙을 이용해서 적절한 품사와 단어를 추정한다[14]. 또한 [16]에서는 미등록어에 대한 음절 패턴을 이용해서 단어와 품사를 추정하고, 품사 부착 시스템에 의해서 단어를 분리하고, 품사를 결정한다. 더구나 ETRI 말뭉치에서 미등록어가 차지하는 비율은 3.45%(<표 1>)에 달한다[4]. 둘째, 형태소 분석기 내에 복합명사를 처리하기 위한 기능이 포함되어야 한다. 복합명사는 단순한 어휘정보와 구문정보만으로 분리할 수 없으며 본질적으로 의미적이며 문맥적인 정보가 요구된다. 예를 들어, 복합명사 "대학생선교회"가 "대학 + 생선 + 교회"와 "대학생 + 선교회"로 분석될 수 있으나, 전자는 어휘적이고 구문적으로는 정당한 해석이나 의미적으로는 적합하지 않은 해석이다. 셋째, 단어 분리만으로도 최근에 활발하게 연구되고 있는 정보검색, 추출요약, 정보추출 등과 같은 응용분야에 두루 사용될 수 있다.

4) 이 자료는 ETRI 말뭉치[17]로부터 얻었다. 학습 말뭉치와 시험 말뭉치를 약 10대 1의 비율로 나누었다. 영역이나 장르에 따라 미등록어의 분포가 다르기 때문에 각 영역이나 장르에서 골고루 추출하였다. 이와 같은 방법을 사용하지 않을 경우에는 미등록어가 차지하는 비율이 훨씬 더 높을 수 있다. <표 1>에서 미등록 어절이란 미등록 형태소를 포함한 어절 수이다.

〈표 1〉 ETRI 말뭉치에서 미등록어 비율

말뭉치 구분	학 습	시 험
어절 수	229,315	26,073
미등록 어절 수		900
미등록 어절 비율		2.78%
형태소 수	304,657	34,665
미등록 형태소 수		962
미등록 형태소 비율		3.45%

2.4 기계학습 방법

최근에 자연언어 처리 분야에서도 여러 가지 방법의 기계학습 방법들이 이용되고 있다[18-19]. 기계학습 방법을 자연언어 처리에 적용하기 위해서는 일반적으로 다섯 가지의 과정이 필요하다. 첫 번째 과정은 이미 잘 정의된 학습 알고리즘에 주어진 문제를 적용하기 위한 **부호화(coding)** 과정이다. 예문 (1)은 문장 "나는 학교에서 놀았다."에 대한 음절기반 단어 분리 문제를 해결할 수 있도록 부호화한 것이다.

- 나는 학교에서 놀았다.
  나/+ 는/+ sp/sp 학/- 교/+ 에/- 서/+ sp/sp 놀/+ 았/+ 다/+ ./+

여기서 '+'는 단어 경계를 의미하고 '-'는 단어의 경계가 아님을 의미하고, sp는 공백을 의미한다. 두 번째 과정은 **자질 결정(feature selection)** 과정이며, 이 과정은 전문가에 의해서 수동으로 자질을 결정하는 방법과 특정 시스템을 이용해서 자동으로 자질을 결정하는 방법이 있다. 예문 (1)의 경우, 각 음절의 자질이 오른쪽 음절, 현재 음절, 왼쪽 음절이라면, 음절 '교'의 자질값은 ('학', '교', '에')가 된다. 세 번째 과정은 **학습(learning) 과정**이며, 이 과정에서 여러 종류의 학습 방법을 이용할 수 있는데, 자연언어 처리에서는 결정트리[20], 사례기반 학습[21], 은닉 마르코프 모델[22] 등과 같은 방법들이 주로 사용된다. 네 번째 과정은 **분류(classification)** 과정이며, 학습된 모델과 주어진 입력 자질을 이용해서 부류(class)를 결정하는 방법이다. 이 과정은 학습 방법에 따라 달라진다. 마지막으로 **해독(decoding)** 과정이다. 예문 (1)에서 부호화된 문장을 다시 해독하면 "나 + 는 학교 + 에서 놀 + 았 + 다 +."가 된다.

2.4.1 사례기반 학습

사례기반 학습[23]은 유사도기반(similarity-based) 혹은 예제기반(example-based) 학습이라고도 불리며 지도 학습(supervised learning) 방법의 일종이다. 학습 과정은 빠른 검색을 위해서 유사한 예제를 군집화하거나 여러 가지 방법으로 색인하여 적절한 형식으로 사례를 저장한다. 분류 과정에서는 입력이 주어졌을 때, 저장된 사례와 가장 비슷한 사례를 추출하고, 추출된 사례의 부류를 입력의 부류로 결정하는 방법이다. 이 방법은 단어 발음변환, 품사 태깅, 전치사구 부

착, 명사구 추출 등과 같은 자연언어 처리 분야[23]에 두루 사용되고 있다. 이 방법에서 입력 $X=(x_1, x_2, \cdots, x_n)$와 사례 $Y=(y_1, y_2, \cdots, y_n)$은 특별한 의미를 지닌 자질로 구성된 패턴이며, 두 패턴의 유사도 $\triangle(X, Y)$는 식 (1)과 식 (2)와 같이 정의된다.

$$\triangle(X, Y)=\sum_{i=0}^{n} w_i\,\delta(x_i, y_i) \tag{1}$$

$$\delta(x_i, y_i)=\begin{cases}\dfrac{x_i-y_i}{\max_i-\min_i} & \text{if } i\text{번째 자질}=\text{숫자, else}\\ 0 & \text{if } x_i=y_i, \text{ else}\\ 1 & \text{if } x_i\neq y_i\end{cases} \tag{2}$$

여기서 $\max_i$와 $\min_i$는 각각 i번째 자질이 가질 수 있는 최대값과 최소값을 의미하고, w_i는 i번째 자질의 가중치이다. 일반적으로 식 (3)과 같은 이득률(gain ratio)을 주로 사용하지만, 여러 가지 가능한 가중치 방법[21]이 있다.

$$w_i=\frac{H(C)-\sum_{v\in V_i} P(v)\times H(C|v)}{-\sum_{v\in V_i} P(v)\log_2 P(v)} \tag{3}$$

$$H(C)=-\sum_{c\in C} P(c)\log_2 P(c) \tag{4}$$

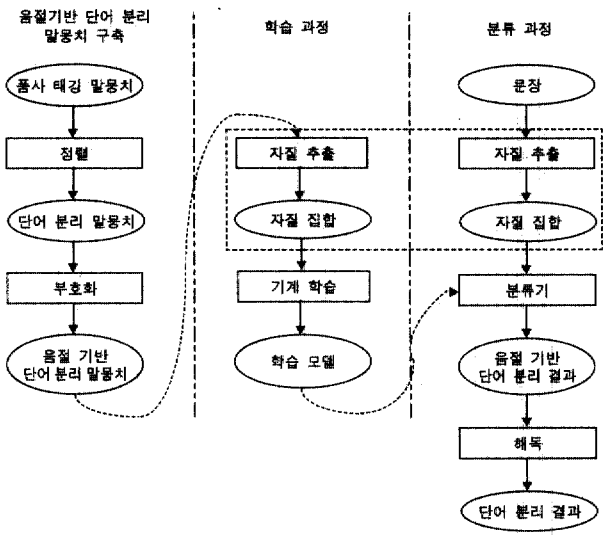
여기서 C는 부류 집합이며, V_i는 i번째 자질에 대한 자질 값 집합이다. 식 (2)에서 기호 자질일 경우, 단순히 일치 여부만 유사도에 반영되고(일치도 측정법(overlap)), 어떤 자질이 특정 부류에 얼마나 기여하는지에 대해서는 유사도에 반영되지 않았다. 특정자질이 특정 부류에 기여하는 정도를 반영하는 유사도 측정법이 식 (5)와 같이 정의된다(기여도 차 측정법(value difference)).

$$\delta(x_i, y_i)=\sum_{k=1}^{n} |P(c_k|x_i)-P(c_k|y_i)| \tag{5}$$

이는 자질 내에 어떤 자질값이 특정 부류와 밀접한 관계를 가지고 있을 때, 좋은 결과를 가져오며, 모든 자질값이 모든 부류에 고른 영향을 줄 경우에는 일치도 측정법과 비슷한 결과를 가져온다.

3. 음절 단위의 한국어 단어 분리

본 논문은 사례기반 학습을 이용한 음절 단위의 한국어 단어 분리 모델을 제안한다. 제안된 시스템의 구성은 (그림 1)과 같으며, 크게 말뭉치 구축, 학습, 분류 과정으로 분리된다. 말뭉치 구축은 기존의 품사 부착 말뭉치로부터 음절기반 단어 분리 말뭉치[5]를 구축한다. 학습 과정은 음절기반 단어 분리 말뭉치로부터 자질을 추출하고, 추출된 자질집합을 이용해서 학습을 수행한다. 분류 과정은 주어진 문장으로부터 자질을 추출하고, 추출된 자질집합과 학습 과정에서 학습된 모델을 이용해서 각 음절 단위로 단어 분리를 위한 범주를 결정한다. 그리고 나서 음절 단위의 부호화된 결과를 해독해서 단어분리 결과를 출력한다. 각 과정에 대한 상세한 설명은 이하의 절에서 기술한다.



(그림 1) 음절 단위로 단어를 분리하기 위한 시스템 구성

3.1 음절 기반 단어 분리 말뭉치 구축

본 논문은 기계학습의 일종인 사례기반 학습을 이용한 음절 단위의 단어 분리 모델을 제안한다. 사례기반 학습은 지도학습이며, 지도학습은 정답에 해당하는 음절 단위의 단어 분리 정보가 부착된 말뭉치가 필요하다. 이를 위해서 본 논문에서는 품사 부착 말뭉치로부터 음절 단위의 단어 분리 말뭉치를 구축한다.

품사 부착 말뭉치에는 어절에 대한 올바른 형태소 분석 결과가 포함되어 있다. 또한 이 결과에는 형태소 분리 결과도 포함되어 있는데, 형태소 분리 결과에 있는 형태소들을 연접하면(concatenate), 원래의 어절과 일치하지 않은 경우가 자주 발생하기 때문에 형태소 분리 결과를 그대로 이용할 수 없다. 본 논문에서는 음절을 기준으로 어절과 형태소 분리 결과를 정렬하여(align), 단어 분리 말뭉치를 구축하는 정렬 단계와 음절기반 단어 분리 말뭉치를 구축하기 위해서 단어 분리 결과를 음절단위로 부호화하는 부호화 단계로 구성된다. <표 2>는 음절 단위의 단어 분리 말뭉치 구축 과정의 예를 보이고 있다[6]. <표 2>에서 첫 번째 열은 어절이며, 두 번째 열은 형태소 분석의 품사 부착 결과이다. 세 번째 열은 어절(첫 번째 열)과 품사 태깅을 이용해서 음절단위의 단어 분리 결과이다(3.1.1절 참조). 네 번째 열은 음절단위의 단어 분리를 부호화한 것이다(3.1.2절 참조).

5) 음절기반 단어 분리 말뭉치에는 음절을 기준으로 올바른 단어 분리 결과를 저장하고 있으며, 좀더 구체적인 설명은 3.1절에서 기술할 것이다.

6) 품사 태깅에서 품사 태그는 [24]를 참조하시오. 단어 분리와 음절 부호화에 대한 범주는 3.1.2에서 자세히 다룰 것이다.

〈표 2〉 음절 단위의 단어 분리와 음절 부호화의 예

어 절	품 사 태 깅	단 어 분 리	음절 단위의 부호화
경제발전이	경제/nc + 발전/nc + 이/jc	경제/n + 발전/n + 이/j	경/n 제/n + 발/n 전/n + 이/j +
비약적으로	비약/nc + 적/xsn + 으로/jc	비약/n + 적/x + 으로/j	비/n 약/n + 적/x + 으/j 로/j +
이루어지고	이루/pv + 어/ec + 지/px + 고/ec	이루/p + 어/e + 지/p + 고/e	이/p 루/p + 어/e + 지/p + 고/e +
교통이	교통/nc + 이/jc	교통/n + 이/j	교/n 통/n + 이/j +
발달하며	발달/nc + 하/xsv + 며/ec	발달/n + 하/x + 며/e	발/n 달/n + 하/x + 며/e +
정복된	정복/nc + 되/xsv + ㄴ/etm	정복/n + 된/x	정/n 복/n + 된/x +
대륙들이	대륙/nc + 들/xsn + 이/jc	대륙/n + 들/x + 이/j	대/n 륙/n + 들/x + 이/j +
세계시장으로	세계/nc + 시장/nc + 으로/jc	세계/n + 시장/n + 으로/j	세/n 계/n + 시/n 장/n + 으/j 로/j +
편입될에	편입/nc + 되/xsv + ㅁ/etn + 에/jc	편입/n + 될/x + 에/j	편/n 입/n + 될/x + 에/j +
따라	따르/pv + 아/ec	따라/p	따/p 라/p +
엄청난	엄청나/pa + ㄴ/etm	엄청난/p	엄/p 청/p 난/p +
인구이동이	인구/nc + 이동/nc + 이/jc	인구/n + 이동/n + 이/j	인/n 구/n + 이/n 동/n + 이/j +
이뤄졌다.	이루/pv + 어/ec + 지/px + 었/ep + 다/ef + ./s	이뤄/p + 졌/e + 다/e + ./s	이/p 뤄/p + 졌/e + 다/e + ./s +

3.1.1 단어 분리를 위한 정렬

품사 태깅 말뭉치에 포함되어 있는 어절과 형태소 분리 결과를 정렬하기 위해서 최소교정거리(minimum edit distance) 알고리즘[25, 153~156 페이지]을 수정하여 이용한다 (그림 2). (그림 2)에서 grapheme_dist(x, y)는 음절 x와 y의 거리를 구하는 함수이며, 식 (6)과 같이 정의된다.

$$\text{grapheme\_dist}(x,y) = 3-\{\delta(x_{\text{초성}},y_{\text{초성}})+\delta(x_{\text{중성}},y_{\text{중성}}) + \delta(x_{\text{종성}},y_{\text{종성}})\} \quad (6)$$

여기서, 함수 $\delta()$의 값은 x와 y가 일치하면 1이고 그렇지 않으면 0이다. 또한 (그림 2)에서 함수 argmin()는 매개변수의 값을 최소로 하는 색인 (i,j)을 구하는 함수인데, 매개변수의 값이 같다면, 색인의 합 $(i+j)$이 큰 것을 선택한다.

최소교정거리를 이용한 어절과 형태소 분리의 정렬 알고리즘

입력 : 어절 $W=(w_1 w_2 \cdots w_m)$; 형태소 분리의 연접 ; $R=(r_1 r_2 \cdots r_n)$;
출력 : 최적경로 $O=(o_1 o_2 \cdots o_m)$

```
행렬 dist (n+1)×(m+1)와 prev (n+1)×(m+1)을 초기화한다 ;
dist[i,0] = ∞ for all i ;
dist[0,j] = ∞ for all j ;
dist[0,0] = 0 ;
for i from 0 to n do
    for j from 0 to m do
        dist[i,j] = min( dist[i-1,j], dist[i-1,j-1],
                         dist[i,j-1]) + grapheme_dist( r_i, w_j) ;
        prev[i,j] = argmin( dist[i-1,j] dist[i-1, j-1],
                            dist[i, j-1]) ;
    enddo
enddo
i = n ; j = m ;
while i > 0 and j > 0 do
    o_j = i ; (i, j) = prev[i, j] ;
enddo
```

(그림 2) 어절과 형태소 분리를 정렬하기 위한 최소교정거리 알고리즘

(그림 3)은 (그림 2)에서 제안된 정렬 알고리즘을 어절 W = (이, 뤄, 졌, 다, .)와 형태소 분리 R = (이, 루, 어, 지, 었, 다, .)에 적용한 예를 보이고 있다. (그림 3)에서 wt_i는 단어 분리 범주이고, rt_i는 부호화된 음절 범주이며, 자세한 설명은 다음 절에서 다룰 것이다. 그 결과 최적 경로 O는 (1, 2, 3, 6, 7)이며, 이를 이용한 정렬 결과 WT는 (이→이, 뤄→루, 졌→어지었, 다→다, .→.)이고, 단어 분리 결과는 "이뤄/p + 졌/e + 다/e + ./s"이다.

s	7	.	5	∞	12	11	11	11	11	12	**9**
e	6	다	4	∞	9	8	8	8	10	**9**	12
e	3	졌	3	∞	6	5	**5**	**7**	**9**	13	16
p	2	뤄	2	∞	3	**2**	5	8	11	15	18
	1	이	1	∞	**0**	3	6	9	12	15	18
wt_i	o_j	w_j	0	**0**	∞	∞	∞	∞	∞	∞	∞
			j / i	0	1	2	3	4	5	6	7
			r_i		이	루	어	지	었	다	.
			rt_i		pv		ec	px	ep	ef	s

(그림 3) 어절 "이뤄졌다."에 대한 정렬 예

3.1.2 음절 범주와 부호화

본 논문은 음절에 기반하여 단어를 분리하고, 그 단어의 범주를 결정한다. 다시 말하면, 작은 단위(음절)의 범주를 결정하여 더 큰 단위(단어)의 범주를 결정하는 문제이다. 이와 같은 문제를 부호화하는 방법에는 크게 두 종류가 있다. 하나는 [26]에서 제안된 **inside/outside 부호화**이고, 다른 하나는 [27]에서 제안된 **open/close 부호화**이다. 본 논문에서 사용하는 방법은 전자와 매우 비슷하다. [26]에서는 명사구가 연달아 나올 경우에 뒤에 나오는 명사구의 시작 단어에 명사구의 경계를 표시하였다. 본 논문에서는 단어가 분리되는 모든 위치에 단어의 경계를 표시한다. 예를 들어, 어절 "경제발전이"를 단어 분리를 위한 음절단위로 부호화

하면 "경/n 제/n + 발/n 전/n + 이/j + "이며 "제"와 "전" 그리고 "이" 다음에서 단어가 분리된다<표 2>. 본 논문에서 음절 단위의 단어 분리를 위한 음절 범주는 <표 3>과 같고, 음절 범주와 음절 경계 범주로 구성되었으며 각각 10개의 범주 총 20개의 범주로 구성되었다. 각 범주에 대한 구체적인 예를 <표 2>에서 찾아볼 수 있으며, 이 범주는 [24]를 기준으로 만들어졌고 그 대응관계를 <표 3>의 네 번째 열에 기술되어 있다.

〈표 3〉 음절단위의 단어 분리를 위한 음절 범주

	음절 범주	음 절 경계 범주	(ETRI, 1999)에서 대응하는 품사 태그
체 언	n	n+	nc, nb
용 언	p	p+	pa, px, pv
수식언	m	m+	maj, maj, mm
독립언	i	i+	ii
조 사	j	j+	jc, jx, jj, jm
지정사	c	c+	co
어 미	e	e+	ef, ec, etn, etm
외국어	f	f+	f
접 사	x	x+	xp, xsn, xsv, xsm
기 호	s	s+	s

3.2 학 습

본 논문에서는 기계학습 방법으로 사례기반 학습을 이용하며 기본적인 학습 방법과 알고리즘은 이미 2장에서 언급되었다. 본 절에서는 문제에 따라 달라지는 자질집합에 대해서 기술한다.

3.2.1 자질 집합

분류기의 성능은 학습 알고리즘에 커다란 영향을 받지만, 자질 집합에도 크게 영향을 받는다. 자질 집합을 결정하는 방법은 자동 방법과 수동 방법이 있다. 본 논문은 후자의 방법으로 자질 집합을 결정하였다. 음절단위의 단어 분리를 위한 i번째 음절 s_i의 자질집합은 왼쪽 음절 $s_{i-L}\cdots s_{i-2}\cdots s_{i-1}$, 자신의 음절 s_i, 오른쪽 음절 $s_{i+1}s_{i+2}\cdots s_{i+R}$, s_i의 받침 f_i 그리고 이전 음절의 범주 $t_{i-T}\cdots t_{i-2}\cdots t_{i-1}$로 구성된다. 여기서 L, R, T는 각각 왼쪽 음절 문맥의 크기, 오른쪽 음절 문맥의 크기, 이전 범주 문맥의 크기이다. <표 4>는 본 논문에서 사용된 각 자질들에 대한 자질값을 보여주고 있다.

〈표 4〉 자질들의 자질값

자질	자질값
s_i	한글 2 바이트 완성형 문자, _SP_, _BOS_
f_i	한글 종성(ㄱ, ㄴ, ㄷ, …, ㄲ, 싸, …, ㄴㅈ, ㄴㅎ, … 등) V, _N_, _E_, _J_, =
t_i	<표 3>의 음절 범주와 음절 경계 범주, _SP_, _BOS_

<표 4>에서 '_SP_'는 공백을 표현하는 음절이자 음절 범주이고, '_BOS_'는 문장의 경계를 나타내는 음절이자 음절 범주이다. f_i에 대한 자질값으로 사용되는 'V', '_N_', '_E_', '_J_', '='은 각각 종성이 없음, 숫자, 영문자, 한자, 기타를 의미한다. 예를 들어, L과 R이 2이고 T가 1이라고 할 때, <표 2>에 있는 어절 "이뤄졌다."의 자질은 <표 5>와 같다.

〈표 5〉 어절 "이뤄졌다."에 대한 자질 집합

s_{i-2}	s_{i-1}	s_i	s_{i+1}	s_{i+2}	f_i	t_{i-1}
이	_SP_	이	뤄	졌	V	_SP_
_SP_	이	뤄	졌	다	V	p
이	뤄	졌	다	.	싸	p+
뤄	졌	다	.	_BOS_	V	e+
졌	다	.	_BOS_	_BOS_	=	e+

3.3 분 류

본 절에서는 분류 과정에 대해서 기술한다. 분류 과정에서 자질 추출 단계는 학습 단계와 동일하기 때문에 본 절에서 기술되지 않는다. 음절 단위로 결정된 범주로부터 단어 단위로 해독하는 과정에 대해서 기술한다.

3.3.1. 단어 분리를 위한 해독

(그림 4)는 단어 분리를 위한 해독 알고리즘이다. 여기서, 함수 removePlus(st_i)는 st_i에있는 경계부호('+')를 제거하며, 예를 들면, st_i가 "p+"라면 rempvePlus(p+)는 p이다. 어절 "이뤄졌다."의 경우로 예를 들어보자. 음절범주 태깅 결과에서 음절열 S이 (이, 뤄, 졌, 다, .)이고, 음절 범주열 ST가 (p, p+, e+, e+, s)를 단어 분리를 위한 해독 알고리즘에 적용하면, 단어 분리 결과인 단어열 W는 (이뤄, 졌, 다, .)이고 단어 범주열은 WT는 (p, e, e, s)이다. 따라서 최종 단어 분리 결과는 "이뤄/p + 졌/e + 다/e + ./s"이다.

단어 분리를 위한 해독 알고리즘

입력 : 음절열 $S=(s_1, s_2, \cdots, s_n)$; 음절 범주열 $ST=(st_1, st_2, \cdots, st_n)$
출력 : 단어열 $W=(w_1, w_2, \cdots, w_m)$; 단어 범주열 $WT=(wt_1, wt_2, \cdots, wt_m)$

```
j = 1 ; fr = 1 ;
for i from 1 to n do
    if ( st_{i-1} ≠ st_i ) then
        w_j = s_(fr,i-1) ; wt_j = st_{i-1} ; fr = i ; j++ ;
    endif
    if ( st_i =~ /^+$/ ) then
        w_j = s_(fr,i) ; wt_j = removePlus(st_i) ; fr = i+1 ; j++ ;
    endif
enddo
w_j = s_(fr,n) ; wt_j = st_n ;
```

(그림 4) 음절 범주로부터 단어를 분리하기 위한 해독 알고리즘

4. 실험 및 평가

4.1 실험 환경

본 논문의 학습 및 시험에 사용될 말뭉치는 ETRI 말뭉

치[17]와 KAIST 말뭉치[28]을 사용하였으며 이들 말뭉치의 통계치는 <표 6>과 같다. <표 6>에서 어절 수는 공백 단위의 단어 수이고, 형태소 수는 품사 부착 말뭉치와 단어 분리 말뭉치(2장 참조)로부터 각각 구했으며, <표 6>에서 알 수 있듯이 일반적으로 품사 부착 말뭉치보다 단어 분리 말뭉치의 형태소 수가 적다(2장 참조). 어절당 평균 형태소 수는 단어 분리 말뭉치의 경우에 해당하는 경우이다. 음절 수는 한글의 경우는 문자 수이고, 영어와 숫자일 경우에는 연속된 문자를 하나로 한 수이다.

〈표 6〉 학습 및 시험 말뭉치의 통계치

말 뭉 치	ETRI			KAIST		
	학 습	시 험	합 계	학 습	시 험	합 계
어절 수	229,315	26,073	255,388	157,938	17,588	175,526
형태소 수(품사 부착)	304,657	34,665	339,322	183,498	20,370	203,868
형태소 수(단어 분리)	270,404	30,746	301,150	165,451	18,465	183,916
어절당 평균 형태소 수	1.18	1.18	1.18	1.05	1.05	1.05
음절 수	731,487	82,847	814,334	483,168	53,823	536,991
형태소당 음절 수	2.70	2.69	2.70	2.92	2.91	2.92

본 논문에서 성능평가의 측도는 음절 단위의 범주 결정의 정확률(accuracy) A, 단어 분리 정확률(precision) P과 재현율(recall) R 그리고 F측도(f-measure) F를 사용하였으며, 이들은 각각 식 (7)~식(10)과 같이 정의된다.

$$A=\frac{N_A^s}{N_S^s}\times 100 \tag{7}$$

$$P=\frac{N_A^w}{N_S^w}\times 100 \tag{8}$$

$$R=\frac{N_R^w}{N_C^w}\times 100 \tag{9}$$

$$F=\frac{2PR}{P+R} \tag{10}$$

여기서 N_S^s와 N_S^w는 각각 시스템이 제시한 음절 수와 단어 수이고, $N_A^s(N_A^w)$은 $N_S^s(N_S^w)$ 중에서 정답에 속하는 음절 수(단어 수)이며 N_C^w는 평가 말뭉치에 속한 단어 수이다. 음절 수의 경우는 평가 말뭉치의 음절 수와 시스템이 제시한 음절 수가 같다. 따라서 음절에 대해서는 정확률과 재현율을 따로 구별하지 않는다.

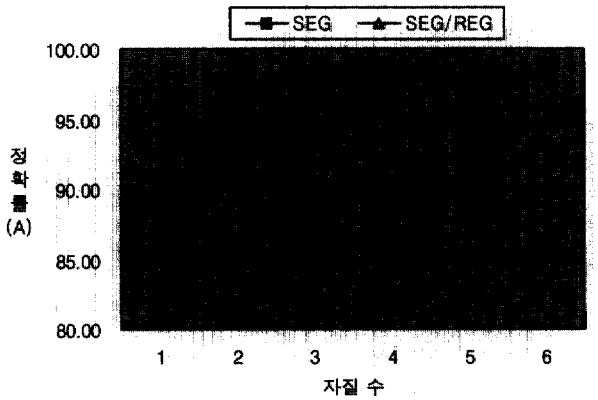
본 논문에서 사례기반 학습을 위해서 사용된 기계학습 도구는 TiMBL(Tilburg Memory Based Learner, version 4.2)[21]를 사용하며, 이하의 실험에서 특별한 언급이 없을 경우에는 이 도구를 이용해서 평가된 것이다.

4.2 자질집합 선택

일반적으로 주어진 자질집합으로부터 분류기에 적합한 자질집합을 선택하는 방법은 크게 평가 함수(filter)와 분류기 성능(wrapper)이 있다[29]. 평가 함수에 의한 자질 선택 방법은 자질에 대한 평가 함수를 정의하고, 그 평가 함수의 값에 따라 그 자질의 포함 여부를 결정한다. 분류기 성능에 의한 자질 선택 방법은 어떤 한 자질이 포함되었을 경우와 그렇지 않을 경우에 대해 분류기의 성능을 비교하여 성능이 좋아질 경우 해당하는 자질을 포함시키는 방법이다. 본 논문에서는 전자의 방법와 후자를 적절히 결합한 방법이다. 먼저, 평가 함수를 식 (4)와 같이 정의하고, 각 자질에 대한 평가 함수의 값에 따라 자질의 순위를 결정한다. 이 순위는 각 자질이 분류기에 포함될 순위이며, 해당 자질이 포함되어 분류기의 성능이 증가하면 그 자질은 분류기에 포함된다. ETRI 말뭉치를 이용해서 각 자질에 대한 중요도와 그 순위는 <표 7>과 같다.

〈표 7〉 자질 선택을 위한 자질 중요도와 그 순위

자 질	p_{i-2}	p_{i-1}	p_i	p_{i+1}	p_{i+2}	f_i	t_{i-2}	t_{i-1}
SEG	0.0114	0.0291	0.0747	0.0574	0.0307	0.1518	0.0470	0.0870
순 위	8	7	3	4	6	1	5	2
SEG/REG	0.0973	0.1842	0.3239	0.2272	0.1272	0.4489	0.2153	0.5078
순 위	8	6	3	4	7	2	5	1



(그림 5) 자질 수에 따른 정확률(A)의 변화

<표 7>의 첫 번째 열에서 SEG와 SEG/REG는 각각 단어 분리 문제와 단어 분리 및 범주 결정 문제[7]의 경우, 각 자질에 대한 평가 함수의 값이다. 본 논문에서 자질 집합을 결정하기 위해서 초기 자질 집합으로 $\{f_i, t_{i-1}, w_i\}$를 사용하였다. 자질은 이득률에 따라서 하나씩 증가시키면서 정확률(A)을 관찰하여 보았다(그림 5). 자질 수가 4와 5개라는 의미는 두 문제 모두에 대해서 자질 집합은 각각 $\{f_i, t_{i-1}, p_i, p_{i+1}\}$이고 $\{f_i, t_{i-2}, t_{i-1}, p_i, p_{i+1}\}$이다. 단어 분리 및

7) 단어 분리 및 범주 결정 문제(SEG/REG)는 주어진 문장으로부터 단어를 추출하고, 각 단어의 범주를 결정하는 문제이고, 단어 분리(SEG)는 주어진 문장으로부터 단어만 추출하는 문제이다.

범주 결정(SEG/REG)에 대해서 자질 수가 7일 경우가 최대이고, 단어 분리(SEG)의 경우에는 자질 수 7과 8에 대해서 큰 변화가 없었다. 따라서 본 논문에서는 자질 집합을 $\{p_{i-1}, p_i, p_{i+1}, p_{i+2}, f_i, t_{i-2}, t_{i-1}\}$로 사용하며, 이하의 모든 실험에서 선택된 자질 집합을 사용한다. 시스템의 속도는 자질 집합의 요수 수에 비례한다. 따라서 시스템의 정확률이 다소 떨어지더라도 속도를 높여야 하는 응용에서는 자질 집합으로 $\{p_i, p_{i+1}, p_{i+2}, f_i, t_{i-1}\}$을 사용할 것을 추천한다.

4.3 모델 변수 최적화

단어 분리(SEG)의 경우는 정확률(A)이 99.04%이므로 모델 변수를 조절하는 것은 큰 의미가 없다. 그러나 단어 분리 및 범주 결정(SEG/REG)재의 경우는 98.34%이므로 약간의 개선의 여지가 있다. 모델 변수는 알고리즘(IB1, IGTREE, TRIBL), 유사측도(일치도, 기여차이), 자질 가중치(이득률, 정보이득, 카이제곱) 등이 있다[21]. 본 논문에서는 모델 최적화 방법으로 등고선탐색법(hill-climbing)을 사용한다. 즉 알고리즘을 결정하기 위해서 TiMBL에서 제공하는 초기 모델 변수(default)를 사용하여 성능이 가장 좋은 알고리즘을 선택한다. 왜냐 하면 어떤 도구를 개발하여 초기 설정된 변수들이 여러 영역에서 좋은 성능을 보이기 때문이다. 선택된 알고리즘을 이용해서 유사도를 결정한다. 선택된 알고리즘과 유사도를 이용해서 자질 가중치를 결정한다. <표 8>은 등고선탐색법에 의한 모델 최적화 결과이다.

〈표 8〉 등고선탐색법을 사용한 모델 최적화

모델 변수	모델 변수 값	정확률(A)	비고
알고리즘	IB1	98.34	초기 모델 변수 : 알고리즘 : IB1 유사도 : 일치도 가중치 : 이득률
	IGTREE	98.10	
	TRIBL	98.34	
유사도	중 복	98.34	알고리즘 : IB1
	기여차이	98.56	
자질 가중치	이득률	98.55	알고리즘 : IB1 유사도 : 기여차이
	정보이득	98.69	
	카이제곱	98.70	

4.4 음절단위의 단어 분리의 견고성

본 논문에서 제안된 방법을 견고성(robustness)을 확인하기 위해서 학습 말뭉치와 다른 영역에 적용해 보았다. 이 실험에서 학습을 위해 ETRI 학습 말뭉치를 사용하고 시험을 위해 KAIST 시험 말뭉치를 사용하였다. <표 9>은 ETRI 말뭉치에 대한 KAIST 말뭉치의 미등록어를 분석한 것이다. 미등록어 비율이 12.6%나 되는데, 미등록어 비율이 높은 이유는 두 말뭉치의 단어 분리 규칙이 크게 다르다. 예를 들면, 어절 "분명하다는"는 ETRI 말뭉치에서는 "분명/n + 하/x + 다/e + 는/e"로 해석되고, KAIST 말뭉치에서는 "분명/n + 하/x + 다는/e"으로 해석되므로 ETRI 말뭉치에서는 단어 '다는'이 미등록어이다. 또한 "한반도비핵화선언"과 같은 복합명사에 대해서도 ETRI 말뭉치는 "한반도/n + 비핵화/n+선언/n"으로 해석되며, KAIST 말뭉치에서는 고유명사로 간주하여 "한반도비핵화선언/n"해석된다. 복합명사에 대해서 ETRI 말뭉치에서는 가능한 한 복합명사를 분리한다. 그러나 KAIST 말뭉치에서는 일반 복합명사는 분리하지만, 고유명사는 분리하지 않는다. 이와 같이 단어 분리의 규칙이 다르지만 <표 10>에서 보는 바와 같이 단어 분리 성능은 크게 떨어지지 않았다. 이 결과로부터 본 논문에서 제안된 시스템이 견고성이 높음을 알 수 있다.

〈표 9〉 ETRI 말뭉치에 대한 KAIST 말뭉치의 미등록어 비율

말 뭉 치	ETRI(학습)	KAIST(시험)
어절 수	255,388	175,526
미등록어절 수		22,132
미등록 어절 비율		12.60%
형태소 수	301,150	183,916
미등록 형태소 수		22,754
미등록 형태소 비율		12.37%

〈표 10〉 ETRI 학습 말뭉치에 대한 KAIST 시험 말뭉치의 성능

말 뭉 치	정확률(A)
단어분리(SEG)	97.09%
단어분리 및 범주 결정(SEG/REG)	95.80%

4.5 음절단위의 단어 분리의 성능

<표 11>은 단어 분리에 대한 시스템의 성능을 보이고 있다. 단어 분리 문제(SEG)에서 단어 분리에 대한 정확률(P)은 ETRI 말뭉치와 KAIST 말뭉치에 대해서 각각 97.56%와 97.42%이다. 이 정확률은 매우 높으며, 약간의 후처리 기능을 보강한다면 실용적인 시스템으로 사용할 수 있을 것이다. 더구나 실용적인 시스템에서 문제가 되는 속도이다. 이를 평가하기 위해서 시스템의 속도를 측정해 보았다. 실험에 사용된 CPU는 UltraSPARC-IIi 333MHz이며 111,835음절(약 3만 단어, ETRI 시험 말뭉치)에 대해서 35초가 소요되었다. 이 정도의 속도라면 충분히 실용적으로 사용할 수 있을 것으로 생각된다. 단어 분리 및 범주 결정문제(SEG/REG)에서

〈표 11〉 단어 분리의 성능평가

말 뭉 치		ETRI	KAIST
단어분리 (SEG)	A	99.00	99.02
	P	97.56	97.42
	R	97.49	97.44
	F	97.53	97.43
단어분리 및 범주 결정 (SEG/REG)	A	98.70	98.96
	P	96.87	97.45
	R	97.45	97.97
	F	97.16	97.71

도 단어 분리 및 범주 결정에 대한 정확률(P)은 ETRI 말뭉치와 KAIST 말뭉치에 대해서 각각 96.87%와 97.45%로 충분히 실용적으로 사용될 수 있을 것으로 생각된다.

4.6 단어 분리의 응용 및 고찰

단어 분리는 색인어 추출, 복합명사 분리, 개체명 인식, 형태소 분석 등 여러 응용 분야에 사용될 수 있다. 앞에서도 언급했듯이 정보검색의 문서 색인으로 사용될 수 있다. 일반적으로 명사(체언)의 경우는 음절 단위로 분리해도 정확한 단어(혹은 형태소)를 찾을 수 있다. 그러나, 대화체의 경우에는 체언의 경우에도 음절 단위로 분리하면 단어를 정확하게 찾을 수 없다. 예를 들면 "바둑인 잠을 잤고 고양인 밥을 먹었다."에서 "바둑인"과 "고양인"을 해석하면 "바둑이 + 는"과 "고양이 + 는"이 되므로 음절 단위로 단어를 분리할 경우에는 단어 "바둑이"와 "고양이"를 정확하게 찾을 수 없다. 단어 분리에서 복합명사는 기본적으로 분리되기 때문에 정보검색 분야에서 색인어 추출을 위해서 필요한 복합명사 분리 모듈을 사용하지 않아도 된다. 최근 정보추출을 위해서 개체명 인식 문제를 다루고 있다. 개체명은 고유명사(인명, 지명, 사건명 등)와 수식표현(절대적인 시간 표현, 기간, 등)을 대상으로 한다[30]. 개체명을 인식하는 대부분의 시스템은 품사 부착 시스템[31]을 사용하는데, 앞에서 언급했듯이 품사 부착 시스템은 미등록어에 대한 문제를 가지고 있으며, 미등록어의 대부분은 고유명사이다. 본 논문에서 단어 분리 시스템은 앞 절에서 언급했듯이 어느 정도 견고하므로 품사 부착 시스템 대신에 이용될 수 있을 것이다. 일반적으로 단어 분리는 형태소 분석기의 한 모듈로 사용할 수 있다[14]. 본 논문에서 제안된 단어 분리 모듈을 형태소 분석기의 단어 분리 모듈로 사용함으로써 시스템의 속도와 형태소 분석의 모호성을 줄일 수 있을 것으로 기대된다.

본 논문은 최소교정거리 알고리즘을 활용하여 음절기반 단어분리 말뭉치를 구축하는 방법을 제안하였으며, 이 방법에 의해서 자동적인 방법으로 음절기반 단어분리 말뭉치를 구축할 수 있었다. 또한 실험을 통해서 이 방법이 매우 유용함을 알 수 있었다. 사례기반 학습 방법을 포함한 기계학습 방법을 이용하여 단어 분리 방법을 제안하였는데, 이 방법은 학습말뭉치가 충분할 경우, 사전과 같은 대량의 언어정보를 구축하지 않고도 단어 분리가 가능하게 되었다. 그러나 사전 정보와 같은 언어 정보를 전혀 이용하지 않는 것은 아니다. 왜냐하면 학습 말뭉치에는 사전 정보뿐 아니라 특정 단어의 용례를 비롯하여 함께 자주 사용되는 연어 정보 등 매우 다양한 정보들이 포함되어 있으므로 기계학습 방법에 의해서 추출된 규칙은 이와 같은 정보들을 일반화된 총체적인 언어정보에 해당한다. 이렇게 함으로써 학습되지 않은 입력 자료에 대해서도 잘 적응할 수 있는 능력을 가지고 있다. 또한 단어 분리만으로도 응용이 가능한 정보검색이나 정보추출 등의 연구 분야의 전처리 시스템으로 아주 효과적인 시스템을 제공할 수 있을 것으로 기대된다.

5. 결론 및 앞으로의 연구과제

본 논문에서는 사례기반 학습 방법을 이용한 단어 분리 모델을 제안하였다. 일반적으로 한국어에서 단어를 분리하기 위해서 형태소 분석기나 품사 부착 시스템을 사용한다. 이들 시스템은 매우 복잡하고, 이를 구축하기 위해서는 복잡한 형태·구문적인 언어 정보와 방대한 사전정보가 요구된다. 본 논문에서 형태소 분석기와 품사 부착 시스템과 같은 언어 처리 시스템을 사용하지 않고 단어 분리 방법을 제안하였으며 음절단위로 단어가 분리된다. 본 논문에서 제안된 방법의 장점을 살펴보자. 첫째, 단시간 내에 시스템을 구축할 수 있다. 최근 한국어에서도 품사 부착 말뭉치를 쉽게 구할 수 있으며 많은 기계학습 도구도 공개 영역에서 쉽게 구할 수 있기 때문에 이들을 이용하면 짧은 시간 내에 단어 분리 시스템을 쉽게 구축할 수 있다. 둘째, 간단한 자질만으로도 비교적 높은 정확률을 얻을 수 있었다. 셋째, 비교적 견고한 시스템으로 영역에 큰 영향을 받지 않는다. 그러나, 자질의 수가 늘어나면 속도가 다소 떨어질 수 있고, 음절을 기반으로 단어를 분리하기 때문에 자연언어 처리 시스템에 그대로 적용할 수 없다는 단점을 가지고 있다.

본 논문에서 제안된 시스템을 평가하기 위해서 ETRI 말뭉치와 KAIST 말뭉치를 사용하였다. 단어 분리 문제에서 단어 분리의 정확률(P)과 재현율(R) 모두가 약 97% 이상을 보였으며, 단어 분리 및 범주 결정 문제에서도 단어 분리의 정확률(P)과 재현율이 평균적으로 97% 이상을 보였다. 충분히 실용적으로 사용할 수 있을 것으로 생각된다. 더구나 견고성 면에서도 좋은 결과를 가져왔다.

앞으로 음소 단위의 단어 분리 모델을 연구하여 음절 단위의 단어 분리 모델의 문제점을 개선하고자 한다. 단어 분리에 대한 후처리 시스템에 대한 연구를 수행해서 사례기반 학습 방법의 문제를 개선하고자 한다. 또한 앞에서 언급한 여러 응용분야에 단어 분리 모델을 적용하고자 한다.

참 고 문 헌

[1] 연세대학교 언어정보개발연구원, 연세 한국어사전, 두산동아, 1998.

[2] 김재훈, "가중치망 모델을 이용한 한국어 품사 태깅", 정보과학회논문지, 제25권 제6호, pp.951-959, 1998.

[3] 이상주, 류원호, 김진동, 임해창, "품사태깅을 위한 어휘문맥 의존규칙의 말뭉치기반 중의성주도 학습", 한국정보과학회논문지(B), 제26권 제1호, pp.178-189, 1999.

[4] Brent, M., "An efficient, probabilistically sound algorithm for segmentation and word discovery," *Machine Learning*, Vol.34, pp.71-106, 1999.

[5] Venkatarman, A., "A statistical model for word discovery in transcribed speech," *Computational Linguistics*, Vol.27, No.3, pp.351-372, 2001.

[6] Allan, J., Callan, J., and Croft, B., "INQUERY at TREC-5," *Proceedings of The Fifth Text REtrieval Conference (TREC-5)*, pp.119-132, 1996.

[7] Sproat R., Shih C., Gale W., Chang N., "A stochastic finite-state word-segmentation algorithm for Chinese," *Computational Linguistics*, Vol.22, No.3, pp.377-404, 1996.

[8] Palmer, D. D., "A trainable rule-based algorithm for word segmentation," *Proceedings of ACL-97*, pp.321-328, 1997.

[9] Brill, E., "Transformation-based error-driven learning and natural language processing : A case study in part-of-speech tagging," *Computational Linguistics*, Vol.21, No.4. pp. 543-565, 1995.

[10] Lua, K.-T. and Gan, K.-W., "An application of information theory in Chinese word segmentation," *Computer Processing of Chinese and Oriental Languages*, Vol.8, No.1, pp. 115-124, 1994.

[11] Yao, Y. and Lua, K.-T., "Splitting-merging model for Chinese word tokenization and segmentation," *Natural Language Engineering*, Vol.4, part 4, pp.309-324, 1998.

[12] Teahan, W. J., Wen, Y., McNab, R. J., Witten, I. H., "A compression-based algorithm for Chinese word segmentation," *Computational Linguistics*, Vol.26, No.3, pp.375-393, 2000.

[13] 이준호, 안정수, 박현주, 김명호, "한글 문서의 효과적인 검색을 위한 n-gram 기반의 색인 방법", 정보관리학회지, 제13호 제1호, pp.47-63, 1996.

[14] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 컴퓨터공학과 박사학위논문, 1993.

[15] 신중호, 박혁로, "음절단위 bigram정보를 이용한 한국어 단어 인식모델", 제9회 한글 및 한국어 정보처리 학술대회 발표논문집, pp.255-260, 1997.

[16] Lee, G. G., Cha, J. and Lee, J.-H., "Syllable-pattern-based unknown morpheme segmetation and estimation for hybrid part-of-speech tagging of Korean," *Computational Linguistics*, Vol.28, No.1, pp.53-70, 2002.

[17] 이현아, 이원일, 임선숙, 허은경, 이재성, 차건희, 박재득, "표준안에 따른 품사 부착 말뭉치 구축", 제11회 한글 및 한국어 정보처리 학술대회 및 제1회 형태소 분석기 및 품사태거 평가 워크숍논문집, 전북, pp.40-43, 1999.

[18] Cardie, C. and Mooney, R. J., "Introduction : Machine learning and natural language," *Machine Learning*, Vol.34, nos.1/2/3, pp.5-10, 1999.

[19] Hammerton, J., Osborne, M., Armstrong, S., and Daelemans, W., "Introduction to special issue on machine learning approaches to shallow parsing," *Journal of Machine Learning Research*, Vol.2, pp.551-558, 2002.

[20] Quinlan, J. R., *C4.5 : Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[21] Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A., *TiMBL : Tilburg Memory Based Learner*, version 4.0, Reference Guide, Technical Report 01-04, Induction of Linguistic Knowledge, Tilburg University, 2001.

[22] Rabiner, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, Vol.77, No.2, pp.257-286, 1989.

[23] Daelemans, W., van den Bosch, A., and Zavrel, J., "Forgetting Exceptions is Harmful in Language Learning," *Machine Learning*, Vol.34, No.1-3, pp.11-41, 1999.

[24] ETRI, 품사 태그 부착 말뭉치 구축 지침서, 한국전자통신연구원, 컴퓨터소프트웨어 기술연구소, 지식정보연구부, 1999.

[25] Jurafsky, D. and Martin, J. H., *SPEECH and LANGUAGE PROCESSING : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, 2000.

[26] Ramshaw, L. and Marcus, M., "Text chunking using transformation-based learning," *Proceedings of the Third Workshop on Very Large Corpora*, pp.82-94, 1995.

[27] Sekine, S. and Grishman, R. and Shinnou, H., "A decision tree method for finding and classifying names in Japanese texts," *Proceedings of the Sixth Workshop on Very Large Corpora*, 1998.

[28] 김재훈, 김길창, 한국어에서의 품사 부착 말뭉치의 작성 요령 : KAIST 말뭉치, 한국과학기술원, 전산학과, CS-TR-95-99, 1995.

[29] Aha, D. W. and Bankert, R. L., "Feature selection for case-based classification of cloud types : An empirical comparison," *Proceedings of the 1994 AAAI Workshop on case-based reasoning*, pp.106-112, 1994.

[30] Chinchor, N., Brown, E., Ferro, L. and Robinson, P., Named entity recognition task definition, version 1.4, 1999.

[31] 김재훈, 투표 방식의 비지도식 모델을 이용한 개체명 분류, 한국과학기술원 전산학과, 석사학위논문, 2002.

김 재 훈

e-mail : jhoon@mail.hhu.ac.kr

1986년 계명대학교 전자계산학과(학사)

1988년 한국과학기술원 전산학과(공학석사)

1996년 한국과학기술원 전산학과(공학박사)

1988년~1997년 한국전자통신연구원, 선임연구원

1997년~1999년 한국해양대학교 컴퓨터공학과 전임강사

2000년~2002년 한국과학기술원 첨단정보기술연구센터 연구원

2001년~2002년 USC, Information Sciences Institute 방문연구원

1999년~현재 한국해양대학교 컴퓨터공학과 조교수

관심분야 : 자연언어처리, 한국어 정보처리, 정보검색, 정보추출

이 공 주

e-mail : kong-joo@hotmail.com

1992년 서강대학교 전자계산학과(학사)

1994년 한국과학기술원 전산학과(공학석사)

1998년 한국과학기술원 전산학과(공학박사)

1998년~2003년 (주)한국마이크로소프트 연구원

2003년 이화여자대학교 컴퓨터공학과 강의전담교수

관심분야 : 자연언어처리, 자연어인터페이스, 기계번역, 정보검색